Explainable Autonomy through Natural Language

Francisco Javier Chiyah Garcia
Heriot-Watt University, Edinburgh, UK
GI-Dagstuhl Seminar 2019 – Explainable Software for Cyber-Physical Systems
About me

• 5th Year student of MEng in Software Engineering.

• Worked for 6 months at SeeByte (software for underwater vehicles and sensors).

• Main contribution: MIRIAM, a multimodal interface for autonomous underwater vehicles.

• Areas: explainability, NLP, NLG, autonomy, augmented-reality...

• Human-Robot Interaction centred.
Robots and Autonomous Systems

- Increasingly being operated remotely, particularly in hazardous environments (Hastie et al., 2018).
- These can instil less trust (Bainbridge et al., 2008).
- Thus, the interface between operator and autonomous systems is key (Robb et al., 2018).
Transparency

• Robots and autonomous systems are hard to understand for non-experts.
• This lack of transparency of how a robot behaves is reflected in decreased trust and understanding.
• Decreased trust and understanding have negative effects on human-machine cooperation.
• Transparent systems are able to provide explanations.
Trust in Autonomous Systems
Mental Models and Explanations 1

- Mental models strongly impact how and whether systems are used.
- Explanations contribute to building accurate mental models of a system.
- Improving the user’s mental model can provide increased confidence and performance (Le Bras et al., 2018).
- According to (Gregor and Benbasat, 1999; Kulesza et al., 2013), “users will not expend effort to find explanations unless the expected benefit outweighs the mental effort”.

What is it doing? (function)  How does it work? (structure)
Mental Models and Explanations 2

• Lim et al. (2009) showed that:

  • explaining “why” a system behaved in a certain way increased understanding and trust
  • “why not” explanations only increased understanding

• Thus both are important regarding the user’s mental model.

Why did the system do that? (function)

Why didn’t the system do something else? (structure)
MIRIAM: The Multimodal Interface 1

• MIRIAM allows for “on-demand” queries for status and explanations of behaviour.

• Increases the user’s situation awareness.

• Requires little training.

MIRIAM: The Multimodal Interface 2
Explainability

• The conversational agent can:

  • Give information about *what* is happening (*function*)
    e.g. “What is the vehicle doing?”, “What is the battery level of the vehicle?”

  • Explain *why* the vehicles are doing (or did) something (*function*)
    e.g. “Why is the vehicle coming to the surface?”

  • Explain “*why not*” the vehicles did not do an *expected action* (*structure*)
    e.g. “Why is the vehicle not going to Area 1?”
"Why" and "Why not" Explanations

Mental models align easily
Generation Method 1

• ‘Speak-aloud’ method whereby an expert provides rationalisation of the autonomous behaviours.
• Derive a model of autonomy.
• Data received from the vehicles is used to steadily build a knowledge base.
Traversing down provides the trace for “why” or “why not” explanations.
Generation Method 2

• Explanations are generated on-demand from a dynamic database that captures context.

• Template-based NLG.

• Explanations come with a confidence value.

• Example explanation:

  ➢ **User:** Why is the vehicle coming to the surface?
  ➢ **System:** The vehicle is transiting to its safe plane depth (medium confidence).
Explaination Effects

- Investigated the effects of explanations on the user through a study.
- What is the best way to give an explanation?
- “What” and “how” to say it are both important.
- Level of detail of an explanation vs number of autonomy model reasons (soundness vs completeness)
- Are they even “worth it”?
Method Insights

• **Advantages:**
  - Expert knowledge can be transferred easily
  - High-level abstraction
  - User-centred
  - On-demand

• **Disadvantages:**
  - Manual process (‘speak-aloud’)
  - Scalability
  - ML systems may prove hard for an expert to explain
Future Work

• Expand what the conversational agent can understand and process
  ➢ Could we do this automatically?

• Generalisation of the agent
  ➢ Could the agent be useful in other domains/systems?

• Handle uncertainty better
  ➢ What are the best ways to handle it?
Summary

• Understanding *what* a system does and *how* it works is important.
• Transparent systems are able to provide explanations.
• Different types of explanations and effects: “why”, “why not”.
• Users won’t read explanations if they don’t believe it is worth it.
• A conversational agent that gives on-demand information.
ES4CPS

• What is an ES4CPS problem, and/or what is an ES4CPS solution, that I am interested in?
  • What makes a system explainable? Can we achieve a formal definition?
  • Conversational agents as an intuitive way of explaining a system on-demand.

• What is the ES4CPS-related expertise that I can contribute to solving this problem?
  • Human-Robot Interaction.
  • Experience with explanations (why, why not) and their effects.

• What external expertise do I need (possibly from the other participants) in order to work on the problem/solution?
  • Distinct concepts of explainability, discuss what it aims to achieve.
  • Expertise with other explainable systems.
Acknowledgements

- Prof. Helen Hastie
- Dr. David A. Robb
- Dr. Pedro Patron
- Atanas Laskov
References

• Hastie, Helen; Lohan, Katrin Solveig; Chantler, Mike J.; Robb, David A.; Ramamoorthy, Subramanian; Petrick, Ron; Vijayakumar, Sethu; Lane, David: The ORCA Hub: Explainable Offshore Robotics through Intelligent Interfaces. In: Proceedings of Explainable Robotic Systems Workshop, HRI’18. Chicago, IL, USA, 2018.

• Kulesza, Todd; Stumpf, Simone; Burnett, Margaret; Yang, Sherry; Kwan, Irwin; Wong, Weng-Keen: Too much, too little, or just right? Ways explanations impact end users’ mental models. In: 2013 IEEE Symposium on Visual Languages and Human Centric Computing. San Jose, CA, USA, pp. 3–10, Sept 2013.


Thank you for your attention